

PARTIAL LEAST SQUARES (PLS) METHODS: ORIGINS, EVOLUTION AND APPLICATION TO SOCIAL SCIENCES.

Gregoria Mateos-Aparicio Morales¹
Complutense University of Madrid , Spain

1. Introduction.

The first aim of this paper is to set out the origin of PLS methods, the prior knowledge that led the originator of PLS, the Swedish professor Herman Wold, mentor of Karl Jöreskog, the founder of Structural Equation Modeling, to develop PLS techniques.

Next, there is a description of how one of the PLS methods, PLS Regression (PLS-R), emerges as a multivariant technique for dimension reduction in order to eliminate multicollinearity in the set of explanatory variables X of a regression model, reducing the dimension of that set so that the resulting subset of descriptive variables is optimal for predicting the dependent variable Y .

This is followed by an explanation of how another method in the evolution of PLS methods, the PLS-Path Modeling (PLS-PM) method, was developed in parallel. This method was proposed by Herman Wold as an alternative to the method of structural equations based on covariance structure. Structural equation methods operate like systems of simultaneous equations, but the estimation of the coefficients based on analysing the covariance structure is made adjusting the covariance matrix, which entails significant and 'hard' conditions (multivariate normality and large samples).

As an alternative, the PLS (Partial Least Squares) technique is used to estimate in a 'soft' way the coefficients of the system of structural equations with the least squares method, since the solutions obtained are just as reliable as with the covariance-based technique and with fewer constraints, basically in data distribution and sample size. This PLS approach for structural equation modeling is called PLS-Path Modeling.

¹ Gregoria Mateos-Aparicio Morales, Universidad Complutense de Madrid, Facultad de Ciencias Económicas y Empresariales, Campus de Somosaguas, 28223 Pozuelo de Alarcón (Madrid).
Tel. +34 91 394 2900. Fax: +34 91 394 2388. E-mail: goyi7@ccee.ucm.es

Once the basic aim is achieved of showing how PLS-Regression and PLS-Path Modeling appeared, were developed and operate, the other aims are to show algorithms and software proposed by different authors, and how the PLS-Path Modeling method has been implemented in social sciences.

2. Origins of the Partial Least Squares (PLS) Methods.

First, a brief mention must be made of the creator of the PLS methods, Herman Ole Andreas Wold, as the origin of the PLS methods can be traced back to him.

Professor Herman Wold² was born on 25 December 1908 in Skien, Norway; he was the youngest of a family of six brothers and sisters. In 1912, his parents migrated to Sweden; it was there that he received his school education before becoming a student at the University of Stockholm. He obtained his doctoral degree from Stockholm in 1938, with the thesis *A Study in the Analysis of Stationary Time Series*³, after studying under Professor Harald Cramér.

H. Wold was already outstanding in other fields before he began to develop the PLS methods. After obtaining his Doctorate, and before moving to Uppsala, Wold started work on statistical demand analysis, commissioned by the Swedish government. His book *Demand Analysis: A study in econometrics*⁴ (with his research assistant Lars Jureen), became a classic in the field although it was not published in English until 1952 because of the war, even though Sweden remained neutral. Before starting to develop the PLS methods, Wold also made various contributions to utility theory, including his "*Ordinal Preferences or Cardinal Utility?*"⁵ which elicited a very positive response from L.J. Savage and G.L.S. Shackle.

"After a few more years in Stockholm, Wold became the first Professor of statistics at Uppsala University in 1942, where he stayed until 1970. He then moved to Gothenburg as Professor of statistics, staying until 1975, when he moved back to Uppsala. He married Anna-Lisa Arrhenius in 1940, and they had three children: Svante, Maria and Agnes. The three children all became scientists"⁶.

² Gani, J.(ed.) (1982). *The Making of Statisticians*. Springer-Verlag, New York, p. 189

³ Wold, H. (1938). *A Study in the Analysis of Stationary Time Series*. (2nd edn 1954, with appendix by P. Whittle). Almqvist and Wiksell, Stockholm

⁴ Wold, H.O.A; Jureen, L. (1952). *Demand Analysis: A Study in Econometrics*. Stockholm: Almqvist and Wiksell,; also New York: Wiley, 1953.

⁵ Wold, H. (1952) "Ordinal preferences or cardinal utility?". *Econometrica*, vol. 20, No.4, pp 661-664

⁶ Wold, Svante (1997). "Wold, Herman Ole Andreas". In *Leading Personalities in Statistical Sciences. From the seventeenth century to the present*. Johnson, N.L. and Kotz, S. (eds.) Wiley, New York, p.213

On account of his numerous contributions to statistics and econometrics, Herman Wold was elected as a member of the Institute of Mathematical Statistics, the American Statistical Association and the Swedish Academy of Sciences; and was made an honorary member of the Royal Statistical Society, the American Economic Association and the American Academy of Arts and Sciences. He was also a member of the Nobel Economic Science Prize Committee from 1968 to 1980 and gave the presentation speech for the Nobel Economic Science Prize winner Lawrence Klein in 1980. He himself was probably nominated as a candidate for this same Prize, although this information is not revealed for 50 years. H. Wold died on 16 February 1992 in Uppsala, Sweden.

Herman Wold worked on econometric models related to estimation methods for systems of simultaneous equations, but unlike his contemporary colleagues, he always preferred to use methods based on least squares rather than on maximum verisimilitude. Wold studied different estimation techniques using iterative procedures, from which he developed a special method called the Fixed-Point Algorithm⁷. This method uses an iterative ordinary least squares (OLS) algorithm to estimate the coefficients of a system of simultaneous equations. In 1964, after a conference on the Fixed-Point method at the University of North Carolina, Wold decided to modify his algorithm to extend it to the calculation of principal components. As he admitted, this modification, presented in 1966, was based on comments from a conference participant, who gave Wold a clue to calculating principal components using an iterative process⁸. Soon afterwards, Wold also applied the algorithm for calculating Hotelling canonical correlations. Of these procedures, Wynne W. Chin said: "The PLS approach has its origins back in 1966 when Herman Wold presented two iterative procedures using least squares (LS) estimation for single and multicomponent models and for canonical correlation"⁹

These two iterative procedures gave way in 1973 to the NIPALS¹⁰ (Non-linear Iterative Partial Least Squares) algorithm, with which H. Wold showed how to calculate principal components with an iterative sequence of simple regressions

⁷ Wold, H. (1973) "Aspects opératoires des modèles économétriques et sociologiques. Développement actuel de l'estimation 'F.P.' (Point fixe) et de la modélisation NIPALS" (linéarisation par itération de moindres carrés partiels). *Économie Appliquée* N° 2-3-4: 389-421

⁸ Wold, H. (1966). "Estimation of principal components and related models by iterative least squares". In P.R. Krishnaiah (Ed.). *Multivariate Analysis*. (pp.391-420) New York, Academic Press.

⁹ Chin, Wynne W. (1998). "The Partial Least Squares Approach to Structural Equation Modeling", in *Modern Methods for Business Research*, edited by G.A. Marcoulides, Lawrence Erlbaum Associates, New Jersey, p. 297

¹⁰ Wold, H. (1973); "Nonlinear Iterative Partial Least Squares (NIPALS) Modeling: Some Current Developments", in P.R. Krishnaiah [ed.]. *Multivariate Analysis II, Proceedings of an International Symposium on Multivariate Analysis held at Wright State University, Dayton, Ohio, June 19-24, 1972*, pp. 383-407. New York, Academic Press.

using the ordinary least squares method (OLS), and also how to calculate canonical correlations with an iterative sequence of multiple regressions using OLS. Finally, in 1977, the PLS algorithm appeared, an iterative algorithm for finding latent variables, according to Fornell: "In 1977, both of H. Wold's iterative algorithms were followed by the general PLS algorithm, originally called NIPALS for LS (least squares) estimation of path models with latent variables"¹¹, and also according to Geladi: "Herman Wold gives the end of 1977 as the birth date of PLS"¹². How the PLS-Regression method, one of the PLS methods, appeared, will now be shown.

3. The origin of Partial Least Squares Regression (PLS-R) as an alternative to Principal Components Regression (PCR).

First, the focus will be on how, as described in the introduction, the Partial Least Squares Regression (PLS-R or PLS Regression) method emerged in order to remove the problem of multicollinearity in a regression model. When the coefficients of a regression model are to be estimated and there is a relatively large number of explanatory variables X with an extreme dependence relationship between them, multicollinearity exists. The problem of multicollinearity means the regression coefficients can be insignificant to the explained variable and this may cause difficulties in interpreting the regression equation due to erratic coefficient signs.

When this problem appears, the most direct solution is to reduce the dimensionality of X , the set of explanatory variables. The immediate question is then how to carry out this reduction. The answer usually involves finding a set of new variables which are created as a linear combination of the originals in such a way that the problem of multicollinearity is eliminated. The principal components method has been widely used for many years and until recently it was a reference point among dimensionality reduction techniques. The application of the principal components method to regression was usually referred to as Principal Components Regression or PCR.

The PCR performs a principal components analysis on X and these principal components are used as explanatory variables for Y . But the problem of choosing an optimum subset of independent variables, principal components, still exists, since they have been chosen to explain X , but there is no guarantee that the principal components that explain X will be pertinent for explaining Y . PLS Regression finds principal components that explain X and are also the best for

¹¹ Claes Fornell and Jaesung Cha (1994) "Partial Least Squares" in *Advanced Methods of Marketing Research*. Edited by R.P. Bagozzi. Blackwell Publishers. Cambridge, USA, p.52

¹² Geladi, P. (1988) "Notes on the History and Nature of Partial Least Squares (PLS) Modeling", *Journal of Chemometrics* 2: 231-246.

explaining Y. This means that it extends principal components analysis with a regression phase so that X's principal components will explain the covariance between X and Y as far as possible. In other words, PLS Regression attempts to extract latent (non-observable) variables so that they collect most of the variation of the real X (observable) variables in such a way that they may also be used to model the Y response (dependent) variables.

As a result, the PLS-R (Partial Least Squares Regression) technique was developed to avoid the effect of multicollinearity (among other factors) in the estimation of regression parameters. In turn, the PLS-R model seeks to predict dependent variables. In practice, this objective represents an attempt to maximize the explained variance of the said variables (variance of Y explained by the correlation existing between X and Y). Therefore, PLS Regression may be more appropriate for predictive purposes (Chin et al., 2003). Indeed, Wold (1979) affirms that PLS Regression is mainly suited to predictive causal analyses in highly complex situations with poorly developed theoretical understanding. Similarly, Barclay et al. (1995) conclude that PLS-R is generally recommended for predictive research models. In other words, PLS-R is a more prediction-oriented method than PCR, since the latter focuses on reducing the dimensionality of X without taking into account the relationship that exists between X and Y.

The two techniques -one based on Principal Components Regression and one on PLS Regression- are compared in solving the problem of multicollinearity in the estimation of regression parameters. Both PLS-R and PCR aim to reduce dimensionality and to thereby tackle the problems that often occur in sets of explanatory variables which have high multicollinearity. However, the two techniques take different approaches and therefore obtain different results. PCR is to establish the maximum variability or variance of the explanatory variables and PLS-R aims to do the same whilst also taking into account the relationship between X and Y. That is, PLS Regression estimates regression parameters so that the variance of Y explained by the correlation existing between X and Y is maximal, or, equivalently, so that the residual variance of the predictive relationships is minimal.

Nevertheless, despite the above advantages, Partial Least Squares Regression (or PLS Regression) did not develop easily because it was initially difficult to position it within a statistical context and this slowed down its application. For this reason, it is a good idea to review its history. As explained in the previous section, the PLS method was developed by Herman Wold. In 1973, with the algorithm NIPALS, Wold showed how to calculate principal components by means of an iterative sequence of simple ordinary least squares (OLS) regressions, as well as how to calculate

canonical correlations with an iterative sequence of multiple OLS regressions. Indeed, Herman Wold continued to apply the technique to new problems and fields. In the 1980s, research interests in PLS shifted from social sciences to applications in chemistry into what is now known as chemometrics (application of statistical methods to chemical data). The person responsible for this transition was Svante Wold, the son of Herman Wold. In 1983, Svante Wold together with Harald Martens adapted NIPALS to solve the problem of multicollinearity in linear regression models. They developed yet another branch of the PLS techniques in analytical chemistry known as PLS Regression (PLS-R)¹³. Besides providing a solution to the problem of multicollinearity in regression models, the PLS technique also solves the problem that arises when the number of individuals is less than the number of variables and the effect that this has on the estimation of regression coefficients. This gives an idea of the potential of this method in situations with small samples. PLS is a powerful analytical tool due to its minimum demands in terms of measurement scales, sample size and residual distribution. PLS does not need the data to come from normal or known distributions (Falk & Miller, 1992).

4. Structural Equation Modeling based on PLS as an alternative to Structural Equation Modeling based on covariance.

As mentioned in the introduction, in parallel to the PLS-R (Partial Least Square Regression) method, another branch of PLS methods was developed, Structural Equation Modeling based on PLS, later called PLS-Path Modeling (PLS-PM). Herman Wold proposed Structural Equation Modeling based on PLS as an alternative to the Jöreskog method, Structural Equation Modeling based on covariance, known as Structural Equation Modeling (SEM). This latter method estimates coefficients of the system of equations adjusting the covariance matrix, i.e. achieving the best fit of the theoretical covariance matrix, deduced from the model, and the initial empirical covariance matrix, i.e. $\Sigma(\Theta) - \hat{\Sigma}$. But, as mentioned in the introduction, this model requires 'hard' conditions of multivariate normality and large samples. On the other hand, the PLS (Partial Least Squares) technique estimates the coefficients of the system of structural equations using the partial least squares method ('soft' form), which avoids restrictions on the distribution of data and on the large sample size.

In 1970 Karl Jöreskog was invited to a conference in Madison, Wisconsin. It was at this conference that Jöreskog presented the first formulation of the Covariance Structure Analysis (CSA) for estimating a *linear structural equation system* which

¹³ Valencia, J.L.; Diaz-Llanos, F.J. *Regresión PLS en las Ciencias Experimentales*. Editorial Complutense, Madrid, 2003, p.3

came to be known later as LISREL¹⁴. The papers of the conference were published in Goldberger and Duncan's¹⁵ *Structural Equations Models in the Social Sciences* which included Jöreskog's paper¹⁶ in which he unified factor analysis, analysis of covariance structures, and linear structural equations modeling in a single general model.¹⁷

Wold disagreed with the *hard modeling* that Jöreskog proposed for models with latent variables and paths. Jöreskog's approach imposed strong hypotheses on data distribution and required a high number of cases or a large sample size. However, the Wold approach was much more agile and the practical absence of data distribution hypotheses was combined with the advantage that a reduced number of cases was sufficient in order to use the algorithm. It was for this reason that the Wold approach was called *soft modeling*. Both approaches, *hard modeling* and *soft modeling*, to path models or structural equation models were compared by the two in 1982 in "*Soft Modeling: The Basic Design and Some Extensions*" in *Systems under Indirect Observation - Causality Structure Prediction*. The main conclusion of this comparison was that, when certain changes were made to Jöreskog's LISREL algorithm, correlation existed between the estimates provided by the two approaches, although the original algorithm obtained more robust estimates with the PLS method.

Some years after this publication and given the considerable dissemination of the PLS approach, Harald Martens suggested using the term PLS-Path Modeling¹⁸, PLS Structural Equation Modeling, to refer to the PLS approach to structural equation models, in order to differentiate PLS-Path Modeling from PLS-Regression.

Herman Wold presented his "soft model basic design" (Wold, 1982) for the PLS estimation algorithm as an alternative to LISREL (Linear Structural Relations), avoiding many of the restrictive hypotheses of the Covariance Structure Analysis (CSA), i.e. multivariate normality and large samples (hard modeling). Despite the flexibility advantages claimed over Covariance Structure Analysis, PLS-Path Modeling was not extensively used in econometrics or in social sciences. Even when

¹⁴ Lee M. Wolfle(2003) "The Introduction of Path Analysis to the Social Sciences, and Some Emergent Themes: An Annotated Bibliography". *Structural Equation Modeling*, 10 (1), p.2

¹⁵ Goldberger, A. S., & Duncan, O. D. (1973). *Structural equation models in the social sciences*. New York. Seminar.

¹⁶ Jöreskog, K.G. (1973) "A General Method for Estimating a Linear Structural Equation System". In: *Structural Equation Models in Social Sciences*, 85-112. A.S. Goldberger and O.D. Duncan (Eds). London: Academic Press.

¹⁷ Mulaik, S.A. (1986) "Factor Analysis and Psychometrika: Major Developments" *Psychometrika*, Vol. 51, No. 1, p.30

¹⁸ M. Tenenhaus et al. (2005). "PLS Path Modeling" *Computational Statistics & Data Analysis* 48, p.160

both approaches were developed practically at the same time, their subsequent evolution was far from parallel. The main reason for the divergence between both techniques is related to software availability. The Covariance Structure Analysis (CSA) was provided with the LISREL program in the early 1970s. However, PLS-Path Modeling software was not available for many years until Lohmöller's *LVPLS ver1.6* program appeared in 1984.

5. PLS Regression and PLS-Path Modeling

There currently exists a great deal of confusion surrounding PLS, both regarding the author (father or son) and the approach (PLS Regression or PLS-Path Modeling). For this reason, these approaches will now be reviewed to understand their relationship, differences and similarities. Both methods will first be briefly recalled for a better understanding, but their operation will then be described in more detail.

PLS Regression (PLS-R), is a multivariable technique for dimension reduction, used to reduce the number of explanatory variables in a regression problem, with the aim of removing multicollinearity from that set of explanatory variables X , and also so that the subset of explanatory variables obtained will be optimal for predicting the dependent variable Y . The following example¹⁹ illustrates how this model is used : The subjective evaluation of a set of 5 wines can be predicted. The dependent variables predicted for each wine are likeability, and how well it goes with meat, or dessert (as rated by a panel of experts). The explanatory variables would be the price, and the sugar, alcohol and acidity content of each wine. Using the results for the regression coefficients by PLSR, new wines could be assigned to e.g. each course of the meal, given the predictive character of the model and the reduction of dimensionality where there is multicollinearity between the different explanatory variables. In this case the reduced sample size ($n=5$) should be noted, as this is incompatible with a classic regression technique.

PLS-Path Modeling is the PLS approach for models of structural equations, used for estimating the coefficients of a system of structural equations with the partial least squares method, since the solutions obtained are just as reliable as those obtained with the technique based on covariance structure, and it has fewer restrictions, especially on data distribution and sample size. These structural

¹⁹ Abdi, H. (2003). Partial least squares regression (PLS-regression). In M. Lewis-Beck, A. Bryman, T. Futing (Eds): *Encyclopedia for research methods for the social sciences*. Thousand Oaks(CA), Sage: pp. 792-795.

equations work like regression equations and the goal is to estimate their coefficients.

The PLS-Path Modeling methodology assumes that structural models are linear, which means that all regression techniques may be used to estimate structural coefficients. However, the ordinary least squares regression model is most widespread due to its lack of application requirements. As a result, the measure used to evaluate the correctness of the adjustment is the usual one for this kind of model, namely the determination coefficient R^2 (i.e. the quotient between the variability explained by the regression and the total variability). Coefficient estimation can be undertaken using the ordinary least squares method, but if multicollinearity is present in the set of explanatory variables, either in the measurement model or in the structural model, the Partial Least Squares Regression method must be used.

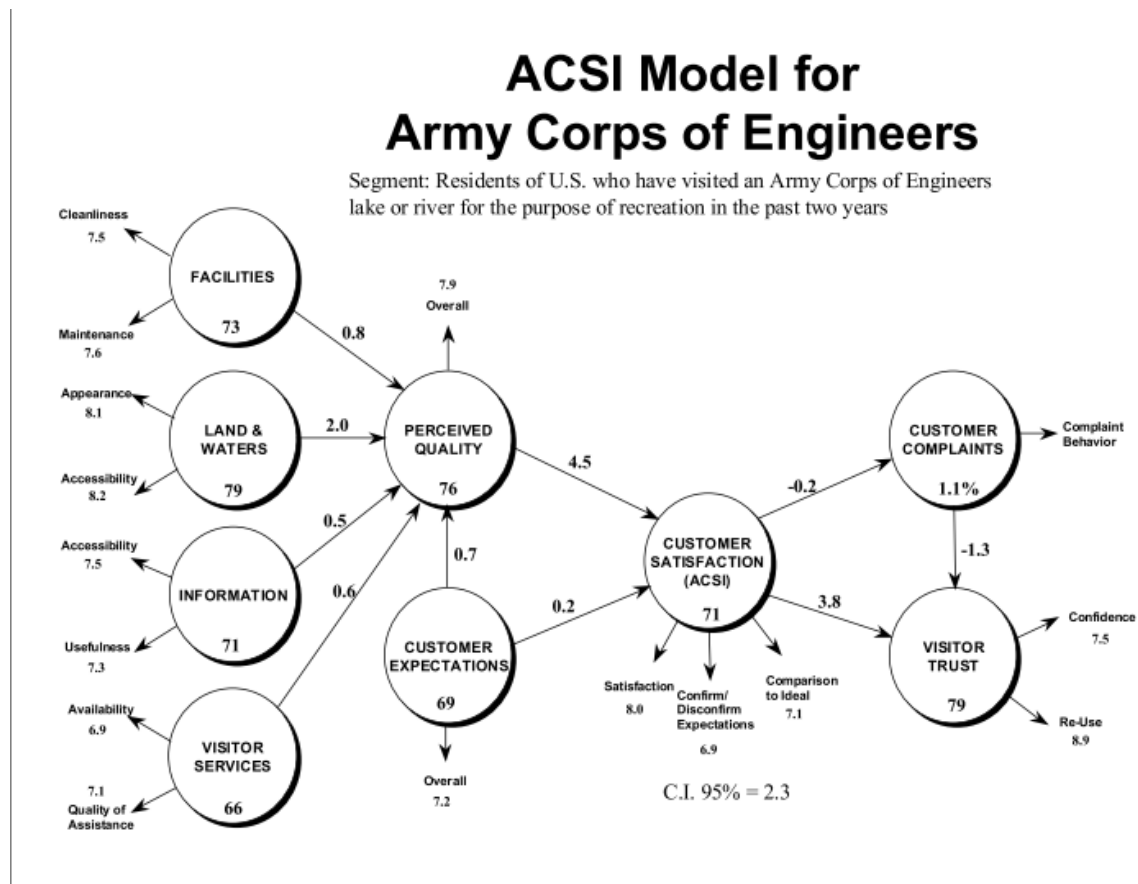
PLS-Path Modeling is the PLS approach for structural equation models, used for estimating the coefficients of a structural equation system with the partial least squares method. The structural equation model combines factorial analysis with path analysis. Factorial analysis leads to what is now called the measurement model, and path analysis to the structural model. Path analysis, which is the basis for structural equation models, is a technique for estimating the unknown parameters of a system of simultaneous equations. It is considered to have been created by the biometrician Sewall Wright who introduced a method in genetics for calculating path coefficients with the publication of his first article in 1918²⁰, in which he included an application of path coefficients to a problem of growth factors. Sewall Wright's path analysis, therefore, fits into PLS-Path Modeling since path analysis determines the coefficients of the equations relating the latent variables in the PLS-Path Modeling structural model. An example of how the PLS-Path Modeling methodology works is shown below²¹

The American Customer Satisfaction Index (ACSI) model is a well established example of the use of PLS-Path Modeling, since it measures ten economic sectors, 43 industries, and more than 200 companies and federal government agencies of the American economy. An application of this model to the Army Corps of Engineers is shown below. A set of nine latent variables is measured through some manifest variables obtained by carrying out telephone interviews. The variables shown in the circles are latent variables and the others are manifest variables. The relationships between the latent variables, shown by arrows, represent the

²⁰ Wright, S. (1918). "On the nature of size factors". *Genetics*, 3, 367-374.

²¹ ACSI; CFI Group; Federal Consulting Group. (2001). Report on *U.S. Army Corps of Engineers* (USACE). Department of Defence

structural model and the relationships shown in the diagram between the latent and manifest variables, represent the measurement model.



A measurement or external model is that part of the complete model that measures the relationship between each latent variable and the associated manifest, or observed, variables. The structural or internal model is that part of the complete model that establishes the relationships between the latent variables. Among the methods for estimation of the complete model, both the sub-models are estimated simultaneously, either using covariance structure analysis, usually termed LISREL (LInear Structural RELationship), which is the algorithm used by SEM (Structural Equation Modeling), or with the Partial Least Squares (PLS) method, which is the estimation method used by PLS-Path Modeling.

When choosing to use the ordinary least squares method or the PLS (Partial Least Squares) method to estimate coefficients, it should be noted that the former does not work well when there is multicollinearity, when there are fewer individuals than variables or when data is missing. In these situations, it is advisable to use the Partial Least Squares (PLS) method. In general, even if there is no multicollinearity in the measurement or structural models of the system of structural equations, the

PLS method can be still used to estimate coefficients, because it provides more precise estimations than those obtained using other methods.

Another advantage of PLS-Path Modeling is that one of the problems that arises with Structural Equation Models (SEM) is their indeterminate nature. In other words, there are too many parameters to estimate for the project's sample size. This is mainly due to the fact that the latent variables are entirely unknown and introduce considerable deficiency into the model when it comes to measuring the relationships that exist between them. PLS resolves this problem easily by simply creating latent variables as the weighted sum of the corresponding manifest variables. This allows two things: firstly, to resolve, as already mentioned, the model's indeterminate nature by obtaining latent variables and, secondly, to analyse the scores obtained for these latent variables. The latter is of great value to companies because each latent variable tends to reflect how they are perceived and companies are always interested in making comparisons with their competitors.

6. Algorithms for PLS Methods.

6.1. Algorithms for PLS Regression

With the algorithms used for performing PLS Regression a distinction has to be made between the PLS1 and PLS2 regression algorithms. The cases must be differentiated where the explanatory variable Y is univariant and where it is multivariant because there are two PLS Regression methods (PLS1 and PLS2) which adapt to said circumstances. In the PLS1 Regression (PLS univariate regression), there is only one variable to explain and there are p explanatory variables, whilst in the PLS2 Regression (PLS multivariate regression) there are q variables to explain ($q > 1$) and p explanatory variables.

PLS1 is the simplest algorithm. With this algorithm the first component is extracted, from which the rest of components are extracted, in such a way that they are uncorrelated (orthogonal). How this algorithm functions will now be described, although not comprehensively, just to show how the partial least squares method works²². The first component is defined thus:

$$t_1 = w_{11}x_1 + w_{12}x_2 + \dots + w_{1p}x_p = \sum_{j=1}^p w_{1j}x_j$$

where x_j are the explanatory variables and y the variable to be explained

²²J.L.Valencia Delfa; J.Díaz-Llanos y Sainz Calleja. *Regresión PLS en las Ciencias Experimentales*. Ed. Complutense, Línea 300, pp:9,13,14

The w_{1j} coefficients are:

$$w_{1j} = \frac{\text{cov}(x_j, y)}{\sqrt{\sum_{j=1}^p \text{cov}^2(x_j, y)}} = \frac{\langle x_j, y \rangle}{\sqrt{\sum_{j=1}^p \langle x_j, y \rangle^2}}, \quad j=1, \dots, p$$

From which it can be deduced that in order to obtain w_{1j} the scalar products $\langle x_j, y \rangle$ must be calculated for each $j = 1, \dots, p$.

Calculating the second component is justified when the single-component model is not adequate, i.e. when the explanatory power of the regression is small and another component is necessary. The second component is denoted by t_2 and it will be a linear combination of the regression residues of the x_j variables on component t_1 instead of on the original variables. In this way, component orthogonality is assured.

To do this, the residue for the single-component regression must be calculated, which will be: $e_1 = y - \hat{y} = y - \hat{\beta}_1 \cdot t_1$, with $\hat{\beta}_1 = \frac{\langle y, t_1 \rangle}{\|t_1\|^2}$.

The second component is obtained thus: $t_2 = w_{21}e_{11} + w_{22}e_{12} + \dots + w_{2p}e_{1p}$

with:
$$w_{2j} = \frac{\text{cov}(e_{1j}, e_1)}{\sqrt{\sum_{j=1}^p \text{cov}^2(e_{1j}, e_1)}} = \frac{\langle e_1, e_{1j} \rangle}{\sqrt{\sum_{j=1}^p \langle e_1, e_{1j} \rangle^2}} \quad \text{para } j=1, \dots, p$$

the residues e_{1j} are calculated by computing the simple regressions of x_j on t_1 , $x_j^* = \hat{\alpha}_j t_1$ ($j=1, \dots, p$), therefore $e_{1j} = x_j - x_j^* = x_j - \hat{\alpha}_j \cdot t_1$, where the estimations of regression coefficients have been calculated thus: $\hat{\alpha}_j = \frac{\langle x_j, t_1 \rangle}{\|t_1\|^2}$

Now with e_1 and e_{1j} , only the scalar products have to be computed $\langle e_1, e_{1j} \rangle$ for each $j = 1, \dots, p$ to be able to calculate component t_2 . To construct the subsequent components the same steps are performed as for the two previous components. This iterative procedure is continued until the number of components to be retained is significant.

The PLS2 algorithm is formulated as an extension of PLS1 for the case when a set of variables $Y = \{y_1, \dots, y_q\}$ has to be explained to relate them with another set of

explanatory variables $X = \{x_1, \dots, x_p\}$. Therefore, a principal component analysis must be performed on a set of variables X with the constraint that these components be as explanatory as possible with respect to the set of variables Y ²³.

6.2. Algorithms for PLS Path-Modeling

The most useful algorithm for PLS-Path Modeling is, without doubt, the NIPALS. As we have already mentioned, this algorithm was created by Herman Wold in 1966 and it offers the great advantage of not requiring the suppression or estimation of surplus or absent observation data in order to use it in the analysis. It allows the parameters of a non-linear model to be estimated by means of a series of simple regressions between the data and certain parameters, hence the name NIPALS (Nonlinear estimation by Iterative Partial Least Squares).

In the PLS-Path Modeling model, PLS supplies the structural equation models with the creation of latent variables from the attributes, created as a linear combination of those attributes or manifest variables. Using SEM, these relationships are studied, but no latent variables are obtained. Therefore, there are two phases in the PLS-Path Modeling algorithm: the estimation of the measurement model and the estimation of the parameters of the structural model - structural parameters-.

Estimating the measurement model produces the weightings of attributes whose linear combination gives the latent variables, and estimating the structural parameters produces an estimation of the coefficients of the relationships between the latent variables. The majority of algorithms calculate the structural coefficients from linear regressions based on the least squares method, but if there are many latent variables or the relationships between them produce a high degree of multicollinearity, then PLS Regression has to be used, since it provides more precise solutions than ordinary least squares regression.

There is no doubt that the clearest explanation of this algorithm is provided by the author Claes Fornell in his work "A Second Generation of Multivariate Analysis, *Vol. I*". Unlike authors such as Löhmoller, Wold or Tenenhaus, Fornell focuses on the applicability of the method rather than the underlying mathematical theory. For this reason, the algorithm offered by Fornell is the most useful approach for most researchers and, as a result, is used as a base to illustrate the PLS-Path Modeling

²³J.L. Valencia Delfa; J.Díaz-Llanos y Sainz Calleja. Regresión PLS en las Ciencias experimentales. Ed. Complutense, Línea 300, p. 29

algorithm. The algorithm that was developed by Claes Fornell for satisfaction studies must also be mentioned. This algorithm adapts the variables (always measured on a scale of 1 to 10) to a scale of 0 to 100. Using this approach, Fornell managed to measure the US economy in terms of satisfaction. In fact, his methodology is still used today for measuring intangible assets.

7. Availability of Software for the PLS Models

There are various software packages that contain the PLS methodology. To avoid making this paper excessively long, this list only aims at providing a reference, without explaining how each of them works, since this is outside the scope of this paper. Among those that include the PLS Regression methodology, the following can be highlighted: SAS, very extended, especially in the chemical and pharmaceutical sectors, where PLS Regression is a very useful tool; and naturally, SIMCA-P, designed by the creators of the technique (Svante Wold and Harald Martens). Other software tools for PLS-Regression can be found at:

http://www.cdpcenter.org/research_scientists/plsr/

The availability of software for PLS-Path Modeling is more limited than of software for structural equation models based on covariance. Despite the greater flexibility achieved with PLS-Path Modeling (the PLS approach for structural equation models) than with the method based on covariance structure, for several years the former was not used much in either econometrics or social sciences. Even though both approximations, PLS-Path Modeling and covariance-based structural equation models, were developed practically at the same time, their evolution was not parallel. The main reason for this divergence is the available software. The method based on analysing covariance structure was supported by Jöreskog's LISREL statistics package from the beginning of the 1970s. Nevertheless, there was no software available for PLS-Path Modeling for years, until Lohmöller's LVPLS 1.6 appeared in 1984, as noted in section 4.

This situation has changed recently, and now researchers can choose from several software options, such as PLSPath (Sellin, 1989), PLS-GUI (Li, 2005), Visual PLS (Fu, 2006a) PLS-Graph (Chin, 2004), SmartPLS (Ringle et al. 2005), SPAD-PLS (Test&Go, 2006) and XLSTAT (Addinsoft, 2008), that are a noticeable improvement, especially for their more intuitive and user-friendly interfaces, on LVPLS (Latent Variable Partial Least Square). The limited dissemination of PLS programmes, together with the difficult and unfriendly LVPLS has led to a reduced dissemination of PLS methodology in relation to the field of Structural Equation Models (SEM). However, a growing number of researchers are beginning to apply this technique

and the distribution of PLS-Graph has contributed to this. Further information is available at: <http://disc-nt.cba.uh.edu/chin/indx.html>

PLS-Graph, developed as part of a project led by Professor Wynn Chin (Chin, 2004) of Houston University; SmartPLS and Visual PLS are graphic versions of Löhmoller's programme with a very easy-to-use and highly intuitive interface, which are fully available on Internet. The latter was designed by Professor Jen Ruei Fu of Kaohsiung University in Taiwan. The main advantage of this tool is that it is entirely free and its reliability has been well demonstrated. Visual-PLS can be downloaded from <http://www2.kuas.edu.tw/prof/fred/vpls/>

8. The PLS-Path Modeling Technique in Social Sciences.

It should be noted that the PLS-Path Modeling technique serves, as it was designed, to reflect the theoretical and empirical conditions of social and behavioural sciences in which we commonly find situations with poorly established theoretical hypotheses and scarce information (Wold, 1979).

The PLS-Path Modeling technique has become highly popular among business management researchers, mainly due to the many advantages it offers in comparison with covariance-based techniques, particularly regarding the demands concerning variable types, sample variable distribution and sample size itself, as mentioned in previous sections.

In the field of social sciences, it must be noted that the PLS-Path Modeling method is an important tool for satisfaction studies. Bagozzi was one of the first authors to propose using PLS techniques for satisfaction studies in the business or marketing world. Others came after him, such as Claes Fornell or Anders Westlund, who developed a specific methodology for satisfaction studies which included the qualitative phase (which obtains the attributes or variables that form the questionnaire) and the interpretation of results. Satisfaction studies currently serve as an important reference framework for market research and customer relations strategies in modern companies.

However, there is a price to pay for the solution offered by PLS in overcoming the indeterminate nature of the model: the parameter estimators are inconsistent and their estimates are biased. In practice, the price to be paid is relatively small since, although consistency may not be achieved in principle, consistency has in fact been proven when the sample size is large enough (in other words, consistency is asymptotic). In SEM models, consistency is ensured if the imposed hypotheses are fulfilled, something which is extremely uncommon in practice.

9. Conclusions

- The aim of this paper, to offer an integrated (although not necessarily comprehensive) view of PLS methods, has been achieved.
- The origins of both PLS Regression and PLS-Path Modeling have been set out and their differences have been analysed.
- The assertion is made that systems of structural equations can be estimated using the PLS method even if no multicollinearity is present, because it produces more accurate estimations.
- PLS-Path Modeling has been compared with Structural Equation Modeling and the advantages of the former over the latter described.
- The PLS1 algorithm has also been stated, to enable a better understanding of the working of the PLS method
- References to available software for the described methods have been included.
- Finally, PLS-Path Modeling has been described as an important research tool in social sciences, especially for satisfaction studies.
- In short, a road-map for the PLS field has been drawn up.

Bibliography

- [1] Abdi, H. (2003). Partial least squares regression (PLS-regression). In M. Lewis-Beck, A. Bryman, T. Futing (Eds): *Encyclopedia for research methods for the social sciences*. Thousand Oaks (CA), Sage, . pp. 792-795.
- [2] Caballero Domínguez, Antonio J. (2006) "SEM vs. PLS: Un enfoque basado en la práctica", IV Congreso de Metodología de Encuestas. Universidad Pública de Navarra.
- [3] Chin, W. W. (1993-2003). *PLS Graph - Version 3.0*. Soft Modeling Inc.
- [4] Chin, W.W. (1998): "The Partial Least Squares Approach to Structural Equation Modeling", in G.A. Marcoulides [ed.]. *Modern Methods for Business Research*, pp. 295-336. Mahwah, NJ: Lawrence Erlbaum Associates, Publisher.
- [5] Chin, W. W. (2000). Frequently Asked Questions - Partial Least Squares & PLSGraph. Home Page.[On-line].
Available: <http://disc-nt.cba.uh.edu/chin/plsfaq.htm>
- [6] Chin, W.W. (2004). PLS-Graph. Version 3.00., Texas (USA). University of Houston
- [7] Falk, R.F. Miller; N.B. (1992). A Primer for Soft Modeling. Akron, Ohio. The University of Akron.

- [8] Fornell, C. (1982) "A Second Generation of Multivariate Analysis: An Overview", in C. Fornell [ed.] *A Second Generation of Multivariate Analysis*, 1, 1-21. New York, Praeger Publishers.
- [9] Fornell, C.; Bookstein, F.L. (1982). "A Comparative Analysis of Two Structural Equation Models: LISREL and PLS Applied to Market Data", in C. Fornell [ed.]. *A Second Generation of Multivariate Analysis*, 1, 289-324. New York, Praeger Publishers.
- [10] Fornell, C. and Cha, J. (1994) "Partial Least Squares" in *Advanced Methods of Marketing Research*. Edited by R.P. Bagozzi. Cambridge, USA, Blackwell Publishers, p.52
- [11] Fu, J.-R. (2006a). *VisualPLS - Partial Least Square (PLS) Regression. An Enhanced GUI for LVPLS (PLS 1.8 PC) Version 1.04*. Taiwan, ROC, National Kaohsiung University of Applied Sciences, .
<http://www2.kuas.edu.tw/prof/fred/vpls/index.html>
- [12] Gani, J.(ed.) (1982). *The Making of Statisticians*. New York, Springer-Verlag, p.189.
- [13] Geladi, P. (1988) "Notes on the History and Nature of Partial Least Squares (PLS) Modeling", *Journal of Chemometrics*, 2: 231-246.
- [14] Jöreskog, K. G. (1973). "A General Method for Estimating a Linear Structural Equation System". In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 85-112). London, Academic Press.
- [15] Jöreskog, K.G.; Wold, H. (1982): *Systems under Indirect Observation - Causality Structure Prediction*. Amsterdam, North Holland Publishing Company.
- [16] Jöreskog Karl G & Sörbom Dag. 1993. LISREL 8 and PRELIS 2. User´s Reference Guide. SSI. Scientific Software International.
- [17] Li, Y. (2005). *PLS-GUI - Graphic User Interface for Partial Least Squares. PLS-PC 1.8. Version 2.0.1 beta*. Columbia, SC, University of South Carolina
- [18] Lohmöller, J.B. (1984). *LVPLS Program Manual. Version 1.6. Latent Variables Path Analysis with Partial Least-Squares Estimation*. Köln Zentralarchiv für Empirische Sozialforschung, Universität zu Köln.
- [19] Mulaik, S.A. (1986). "Factor Analysis and Psychometrika: Major Developments" *Psychometrika*, Vol. 51, No. 1.
- [20] Ringle, C. M., Wende, S., and Will, A. (2005). *SmartPLS - Version 2.0*. Universität Hamburg, Hamburg.
- [21] Sellin, N. (1989). *PLSPATH - Version 3.01. Application Manual*. Hamburg. Universität Hamburg
- [22] Tenenhaus, M. (1998). *La régression PLS*. Paris, Technip.
- [23] M. Tenenhaus et al. (2005). "PLS Path Modeling" *Computational Statistics & Data Analysis*, 48, p.160
- [24] Test&Go (2006). *Spad Version 6.0.0*. Paris, France.

- [25] Valencia, J.L.; Diaz-Llanos, F.J. *Regresión PLS en las Ciencias Experimentales*. Madrid, Editorial Complutense, 2003, p.3
- [26] Wold, H. (1938). *A Study in the Analysis of Stationary Time Series*. (2nd ed 1954, with appendix by P. Whittle). Stockholm, Almqvist and Wiksell,
- [27] Wold, H.O.A; Jureen, L. (1952). *Demand Analysis: A Study in Econometrics*. Stockholm: Almqvist and Wiksell, also New York: Wiley, 1953.
- [28] Wold, H. (1952) "Ordinal preferences or cardinal utility?". *Econometrica*, vol. 20, No.4, pp 661-664
- [29] Wold, H. (1966). "Estimation of principal components and related models by iterative least squares". In P.R. Krishnaiah (Ed.). *Multivariate Analysis.*, New York, Academic Press. pp.391-420
- [30] Wold, H. (1973). "Aspects opératoires des modèles économétriques et sociologiques. Développement actuel de l'estimation "F.P." (Point fixe) et de la modélisation "NIPALS" (linéarisation par itération de moindres carrés partiels)". *Économie Appliquée* N° 2-3-4: 389-421
- [31] Wold, H. (1973). "Nonlinear Iterative Partial Least Squares (NIPALS) Modeling: Some Current Developments", in P.R. Krishnaiah [ed.]. *Multivariate Analysis II*, Proceedings of an International Symposium on Multivariate Analysis held at Wright State University, Dayton, Ohio, June 19-24, 1972, New York: Academic Press, pp. 383-407.
- [32] Wold, H. (1979). *Model Construction and Evaluation when Theoretical Knowledge Is Scarce: An Example of the Use of Partial Least Squares*. Cahiers du Département D'Économétrie. Genève, Faculté des Sciences Économiques et Sociales, Université de Genève.
- [33] Wold, H. (1980): "Soft Modeling: Intermediate between Traditional Model Building and Data Analysis", *Mathematical Statistics*, 6, pp. 333-346.
- [34] Wold, H. (1982). Soft modeling: the basic design and some extensions. In K.G. Jöreskog & H. Wold (Eds.) *Systems under indirect observations: Causality, structure, prediction*. Part 2, Amsterdam, North-Holland, pp.1-54
- [35] Wold H. (1985). "Partial Least Squares", in S. Kotz and N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences (vol. 6)*, New York, Wiley, pp. 581-591.
- [36] Wold, Svante (1997). "Wold, Herman Ole Andreas". In *Leading Personalities in Statistical Sciences. From the seventeenth century to the present*. Johnson, N.L. and Kotz, S. (eds.) Wiley, New York, p.213
- [37] Wold, S. (2001). "Personal memories of the early PLS development". *Chemometrics and Intelligent Laboratory Systems*, 58, 83-84.
- [38] Wolfle, Lee M. (2003) "The Introduction of Path Analysis to the Social Sciences, and Some Emergent Themes: An Annotated Bibliography". *Structural Equation Modeling*, 10 (1), p.2
- [39] Wright, S. (1918). On the nature of size factors. *Genetics*, 3, 367-374.